AD-A037 667 COLORADO STATE UNIV FORT COLLINS DEPT OF MATHEMATICS F/G 12/1
A COMBINED REMES-DIFFERENTIAL CORRECTION ALGORITHM FOR RATIONAL--ETC(U)
NOV 76 E H KAUFMAN, D J LEEMING, G D TAYLOR AF-AFOSR-2878-76
UNCLASSIFIED AFOSR-TR-77-0165 NL

| OF |
AD
A037667

END
DATE
FILMED
4-77

3

A Combined Remes-Differential Correction
Algorithm for Rational Approximation

by

Edwin H. Kaufman, Jr.
Department of Mathematics
Central Michigan University
Mount Pleasant, Michigan 48859

David J. Leeming[1]
Department of Mathematics
University of Victoria
Victoria, B. C., Canada V8W 2Y2

and

G. D. Taylor[2]
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523

[1]Research supported in part by National Council of Canada Grant A8061

[2]Research sponsored by the Air Force Office of Scientific Research,
Air Force Systems Command, USAF,
under Grant No. AFOSR-76-2878

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

ABSTRACT

In this paper a hybrid Remes-differential correction algorithm for computing best uniform rational approximants on a compact subset of the real line is developed. This algorithm differs from the classical multiple exchange Remes algorithm in two crucial aspects. First of all, the solving of a nonlinear system to find a best approximation on a given reference set in each iteration of the Remes algorithm is replaced with the differential correction algorithm to compute the desired best approximation on the reference set. Secondly, the exchange procedure itself has been modified to eliminate the possibility of cycling that can occur in the usual exchange procedure. This second modification is necessary to guarantee the convergence of this algorithm on a finite set without the usual normal and sufficiently dense assumptions that exist in other studies.

AMS(MOS) subject classification numbers: 41A20, 41A50, 65D15.

1. Introduction

This paper is divided into two parts. In the first part we consider X a compact subset of the real line with card(X) $\geq n + 2$. Let C(X) denote the class of all continuous real valued functions defined on X, normed with the uniform norm, i.e., $\|f\| = \max\{|f(x)| : x \in X\}$. Let n be a positive integer and set

$$R_n^0(X) = \{r = 1/p : p \in \Pi_n,\ p(x) > 0 \text{ for all } x \in X\}$$

where Π_n denotes the set of all algebraic polynomials of degree $\leq n$. Note that $R_n^0(X)$ consists of only the positive elements of the set usually denoted by $R_n^0(X)$. In this setting we will give an algorithm for computing the best approximation for positive $f \in C(X)$ from $R_n^0(X)$. We believe this algorithm is the correct analog, for this setting, of the standard multiple exchange Remes algorithm for polynomials. We observe here that if $Y \subset X$, Y is compact and card(Y) $\geq n + 2$, then existence of a best approximant to positive $f \in C(Y)$ from $R_n^0(Y)$ is guaranteed by [5].

This algorithm contains some unique features including the incorporation of the differential correction algorithm [1], [4] to obtain a best approximation at each stage. This insures that the denominator of the best approximation, p_k, on the k^{th} reference set, X_k, will be positive on X_k. If, however, $p_k(x) \leq 0$ for some $x \in X \sim X_k$, we indicate two exchange procedures for selecting the next reference set. Note that in most studies this possibility is ignored by assuming (1) f is normal on some interval [a, b] containing X; (2) X is sufficiently dense in [a, b]; and, (3) X_k is sufficiently close to an alternating set of the best approximation to f on X. We shall also show that using our exchange procedure, there exists a $k_0 > 0$ such that for $k \geq k_0$, p_k must be positive on X. From this point on, our exchange procedure will coincide with the standard multiple exchange procedure and we can therefore guarantee convergence without the above assumptions.

Our procedure could also be used to overcome the difficulty which Dunham [3] has pointed out in William's paper on interpolating rationals [7].

It should be further emphasized that a modified exchange procedure is actually necessary to guarantee the convergence of this algorithm without the assumptions (1)-(3) of above. Indeed, if one attempts to use the standard exchange procedure without regard to the possibility that $p_k \leq 0$ on $X \sim X_k$ may occur (and hoping that $p_k(x) = 0$ for $x \in X \sim X_k$ does not occur to give a divide fault) the usual proof that the error of approximation on the successive reference sets is strictly increasing is false. In fact, examples exist for which the error does not increase strictly and for which the algorithm actually cycles (i.e., $p_k = p_{k+2} = p_{k+4} = \ldots$; $p_{k+1} = p_{k+3} = p_{k+5} = \ldots$; $X_k = X_{k+2} = X_{k+4} = \ldots$; $X_{k+1} = X_{k+3} = X_{k+5} = \ldots$; starting at some k). Using either of the exchange procedures that we give, we are able to prove that the error of approximation on successive reference sets is strictly increasing.

The second part of the paper is devoted to the description of the Remes-Difcor algorithm (the name of our algorithm) for obtaining the best approximation to $f \in C(X)$, X a finite subset of the real line of at least $n + m + 2$ points, by elements of $R_n^m(X)$, $m \geq 0$, $n \geq 0$, where

$$R_n^m(X) = \{r = p/q : p \in \Pi_m,\ q \in \Pi_n,\ q(x) \geq \varepsilon \text{ for all } x \in X\},$$

and the ε is chosen so that a best approximation from $R_n^m(X)$ will also be a best approximation from the larger class that relaxes this requirement to $q(x) > 0$ on X. A proof of the convergence of this algorithm is given, along with a flow chart. Finally, a brief discussion of some numerical results will be given. A complete discussion of the numerical results and comparison with both the Remes algorithm and the differential correction algorithm is planned for in a separate paper.

2. Approximating with $R_n^0(X)$.

Let $f \in C(X) - R_n^0(X)$, with $f > 0$ on X. We first consider the case where X is a finite subset of the real line, with card $(X) \geq n + 2$. For each k, $k = 1, 2, \ldots$, $X_k \subset X$ shall denote a reference set of $n + 2$ or $n + 3$ points and $r_k = 1/p_k \in R_n^0(X_k)$ will denote the best approximation to f on X_k from $R_n^0(X_k)$. This best approximation, r_k, is obtained by using the differential correction algorithm applied to the point set X_k. There are three advantages to finding r_k via the differential correction algorithm rather than via solving a nonlinear system of equations: a solution is guaranteed, we are assured that $p_k > 0$ on X_k, and no extra complications will arise if X_k has $n + 3$ points. After computing r_k, if X_k has $n + 3$ points we delete one point of X_k to get a new set Y_k of $n + 2$ points, taking care that $f - r_k$ alternates on Y_k. If X_k consists of $n + 2$ points, then we set $Y_k = X_k$. Set $e_k = \max\{|f(x) - r_k(x)| : x \in Y_k\}$, $Z_k = \{x \in X: \; p_k(x) > 0\}$ and consider the following two exchange procedures for constructing the next reference set X_{k+1}:

Exchange I: (The positive exchange) If r_k is not the best approximation to f on Z_k from $R_n^0(Z_k)$, X_{k+1} is constructed from Y_k by doing an ordinary Remes multiple exchange on the points of Z_k. If r_k is the best approximation to f on Z_k from $R_n^0(Z_k)$ then the algorithm terminates if $Z_k = X$. If $Z_k \neq X$ then $y \in X$ satisfying $p_k(y) = \min\{p_k(x) : x \in X\}$ is found and X_{k+1} is defined to be $X_{k+1} = \{y\} \cup \{Y_k\}$. Note that in this case we have the $p_k(y) \leq 0$ and X_{k+1} consists of $n + 3$ distinct points of X.

Exchange II. (The negative exchange) In this exchange procedure, the algorithm first does a standard Remes multiple exchange on the point set Z_k with respect to $f - r_k$ and Y_k getting $W_k \subset Z_k$, where W_k consists of $n + 2$ points on which $f - r_k$ alternates in sign, $|f(w) - r_k(w)| \geq e_k$ for all $w \in W_k$ and $\max\{|f(w) - r_k(w)| : w \in W_k\} = \max\{|f(x) - r_k(x)| : x \in Z_k\}$. If $W_k = Y_k$ and

$Z_k = Y_k$ and $Z_k = X$ then the algorithm terminates as r_k is the desired best approximation to f on X. If this does not happen then X_{k+1} is defined to be $W_k \cup \{y\}$ if $Z_k \neq X$ where y satisfies $p_k(y) = \min\{p_k(x) : x \in X\} \leq 0$ and W_k if $Z_k = X$.

Note that this exchange procedure differs from the first one in that whenever $Z_k \neq X$ an additional point were p_k takes on its minimum is added to the reference set. In the first exchange procedure this additional point is added only when r_k is the best approximation to f on Z_k from $R_n^0(Z_k)$. Also, note that whenever $Z_k = X$ both of these procedures coincide with the standard Remes multiple exchange procedure. For both of these exchange procedures the following theorem holds. (The set $X_1 \subset X$ is chosen so that it has n + 2 points and $e_1 > 0$.)

<u>THEOREM 1</u>. If X is finite and the algorithm described above using either of the two exchange procedures is applied, then $\{e_k\}$ is strictly increasing. Furthermore, the algorithm eventually terminates at a best approximation to f on X from $R_n^0(X)$.

<u>Proof</u>: To show that $e_k < e_{k+1}$ for all k one must consider two cases. The first is when X_{k+1} is constructed only from points of Z_k. In this case p_k and p_{k+1} are both positive on X_{k+1} and a standard de La Vallee Poussin type of argument (zero counting) shows that $e_k < e_{k+1}$ since p_{k+1} is best on X_{k+1} and $p_k \neq p_{k+1}$. In the case that $X_{k+1} = W_k \cup \{y\}$ where $W_k = Y_k$ or W_k is the result of a standard Remes multiple exchange on the points Z_k with respect to $f - r_k$ and Y_k, and $y \in X$ satisfies $p_k(y) = \min\{p_k(x) : x \in X\} \leq 0$, there are two subcases to be considered. The first is when $|f(y) - r_{k+1}(y)| < e_{k+1}$ so that $f - r_{k+1}$ alternates on W_k with error e_{k+1}. Since p_k is also positive on W_k and $|f - p_k| \geq e_k$ on W_k we must have that $e_{k+1} > \min\{|f(z) - r_k(z)| : z \in W_k\} \geq e_k$

by the same de La Vallee Poussin type of argument. Finally, if $|f(y) - r_k(y)| = e_{k+1}$ and $f - r_{k+1}$ alternates on $Y_{k+1} \subset X_{k+1}$ where $Y_{k+1} \neq W_k$, then we must have that $\max\{|f(z) - r_{k+1}(z)| : z \in W_k\} = e_{k+1}$. Also, $f - r_k$ alternates in sign on W_k with $|f - r_k| \geq e_k$ on W_k. Thus, by zero counting we must once again have that $\max\{|f(z) - r_{k+1}(z)| : z \in W_k\} > \min\{|f(z) - r_k(z)| : z \in W_k\}$ since $p_k \not\equiv p_{k+1}$ implying that $e_k < e_{k+1}$. (For a more careful treatment of the de La Vallee Poussin type of argument see the proof of Lemma 2 later in the paper.) The rest of the theorem now follows since X is finite, and no reference set can occur more than once.

Although, in actual computation one only encounters finite sets, it is of interest to consider the behavior of this algorithm if X is only required to be compact. In the remainder of this section we shall only consider Exchange I (the positive exchange). It can be shown that similar results are true for Exchange II. We first note that in this case the set $Z_k = \{x \in X : p_k(x) > 0\}$ may fail to be compact. If this happens then it may not be possible to carry out the Remes multiple exchange on Z_k with respect to $f - r_k$ and Y_k. Thus, the algorithm must be modified by choosing some $\varepsilon > 0$ and setting $Z_k = \{x \in X : p_k(x) \geq \varepsilon\}$. The elements of the set $G_k = \{x \in X : p_k(x) < \varepsilon\}$ will be called g-poles (generalized poles) of p_k. The number ε should be chosen so that p_k has no g-poles on X_k. Since

$$\left\|\frac{1}{p_k}\right\|_{X_k} \leq \left\|f - r_k\right\|_{X_k} + \|f\|_{X_k} \leq 2\,\|f\|_{X_k} \leq 2\,\|f\|,$$

it suffices to choose any ε with $0 < \varepsilon \leq \frac{1}{2\|f\|}$ For such a choice of ε, the algorithm is defined as above with either of the two exchanges. We now prove that this modified algorithm converges globally and at least linearly.

<u>THEOREM 2</u>. For X a compact subset of [a, b], and $0<\varepsilon<\frac{1}{2\|f\|}$, and $f \in C(X) - R_n^0(X)$, the rational functions r_k generated by the modified algorithm described above have

no g-poles on X for $k \geq$ some k_0 and converge uniformly to the best approximation r^* to f on X according to an inequality of the form $\|r_k - r^*\|_X \leq A\theta^k$, $0 < \theta < 1$, for $k \geq k_0$.

Proof: Since the conclusion follows trivially if the algorithm terminates, we assume that this is not the case. The method of proof is to show that $\{e_k\}_{k=1}^{\infty}$ is increasing and to actually estimate this rate of increase. To prove that $e_k < e_{k+1}$ holds for all k, one simply uses the arguments of Theorem 1. Also, note that $\{e_k\}_{k=1}^{\infty}$ is bounded (otherwise $r \equiv 1$ would be a better approximation than r_k on X_k for some k). Hence, there exists e^* such that $e_k \uparrow e^*$. The remainder of this proof is broken into seven lemmas; the first of these, which proves that the points in Y_k cannot cluster is proved by arguments similar to Wendroff [6, p. 65].

LEMMA 1. There exists $\delta > 0$ such that for every k, if $Y_k = \{x_0^k, \ldots, x_{n+1}^k\}$, then $x_i^k \leq x_{i+1}^k - \delta$ for $i = 0, 1, \ldots, n - 1$.

Proof: Suppose not, then there exist, for some fixed i, subsequences (relabelled as) $\{x_i^k\}$ and $\{x_{i+1}^k\}$ such that $x_i^k \to x_i^*$ and $x_{i+1}^k \to x_i^*$. By passing to further subsequences (relabelling if necessary) we have that $x_j^k \to x_j^*$ for $j=0,\ldots,n+1$ as $k \to \infty$ where $x_i^* = x_{i+1}^*$ and $x_j^* \leq x_{j+1}^*$, $j = 0, 1, \ldots, n - 1$. Thus, on the set $X^* = (x_0^*, \ldots, x_{n+1}^*)$ we can find $p^* \in \Pi_n$ such that $f(x_j^*) - \frac{1}{p^*(x_j^*)} = 0$, $j = 0, 1, \ldots, n + 1$. By continuity, there exists a $\delta > 0$ such that $\left|f(x) - \frac{1}{p^*(x)}\right| < \frac{e_2}{2}$ for $x \in \bigcup_{j=0}^{n+1} (x_j^* - \delta, x_j^* + \delta) \cap X$ where $e_2 > 0$, is the error of the second cycle. Hence for sufficiently large k, we have that $\left|f(x_j^k) - \frac{1}{p^*(x_j^k)}\right| < \frac{e_2}{2}$, $j = 0, 1, \ldots, n + 1$. But this implies that $e_k \leq \frac{e_2}{2}$ since $\frac{1}{p_k}$ is best on $Y_k = \{x_0^k, \ldots, x_{n+1}^k\}$ which contradicts the fact that $e_\nu \uparrow$. ∎

LEMMA 2. Let X be a compact set of real numbers containing at least $m + n + 2$ points, and let $f \in C(\overline{X})$. Suppose $r^* = \frac{p^*}{q^*} \in R_n^m(\overline{X})$ has defect $d = \min(m-\partial p^*, n-\partial q^*)$ and let $N = m + n + 2 - d$. Suppose that $f - r^*$ alternates in sign on $\{x_i\}_{i=1}^N \subset \overline{X}$ where $x_1 < x_2 < \ldots < x_N$, and that $f(x_i) - r^*(x_i) \neq 0$, for $i = 1, \ldots, N$. Then if $r = \frac{p}{q} \in R_n^m(\overline{X})$, $r \not\equiv r^*$ on $\overline{X}$ we have

$$\max_{1 \le i \le N} |f(x_i) - r(x_i)| > \min_{1 \le i \le N} |f(x_i) - r^*(x_i)|.$$

Proof: Suppose $\max_{1 \le i \le N} |f(x_i) - r(x_i)| \le \min_{1 \le i \le N} |f(x_i) - r^*(x_i)|$. Let $\Delta(x) \equiv r(x) - r^*(x) = (f(x) - r^*(x)) - (f(x) - r(x))$, for all $x \in \overline{X}$. Assume (without loss of generality) that $f(x_1) - r^*(x_1) > 0$; then we have $(-1)^i \Delta(x_i) \le 0$, $i = 1, \ldots, N$. Now for all $x \in \overline{X}$,

$$\Delta(x) = \frac{p(x)}{q(x)} - \frac{p^*(x)}{q^*(x)} = \frac{p(x)q^*(x) - p^*(x)q(x)}{q(x)q^*(x)} \equiv \frac{S(x)}{q(x)q^*(x)}$$

so that $(-1)^i S(x_i) \le 0$, $i = 1, \ldots, N$. But $\partial S \le m + n - d = N - 2$ so $S \equiv 0$. Therefore $r \equiv r^*$ on $\overline{X}$ and this contradiction proves the lemma.

LEMMA 3. There exists a constant c such that for every k, if $p_k(x) = p_0^k + p_1^k x + \ldots p_n^k x^n$, then $|p_i^k| \le c$ for $i = 0, 1, \ldots, n$.

Proof: Suppose not. Let $\frac{1}{p_k} = \frac{c_k}{q_k}$ for all k where $\|q_k\|_X = 1$ and $c_k > 0$. Let $I = [a, b]$ be a closed interval with $a \le \min\{x: x \in X\} - \delta$ and $b \ge \max\{x: x \in X\} + \delta$ where δ is the δ of Lemma 1. Note that $c_k \le 2\|f\|$ for all k. Now, if there exists $\eta > 0$ such that $c_k \ge \eta$ for all k then the desired result follows. Thus, let us assume that there exists a subsequence (which we relabel) for which $c_k \to 0$, $q_k \to \overline{q} \in \Pi_n$ uniformly on I with $\|\overline{q}\| = 1$. Let $z_1, \ldots, z_\ell$ be the distinct zeros of $\overline{q}$ in I, and choose non-intersecting intervals $I_1 = (z_1 - \delta_1, z_1 + \delta_1), \ldots, I_\ell = (z_\ell - \delta_1, z_\ell + \delta_1)$

with $0 < \delta_1 < \frac{\delta}{2}$. Let $J = I - \bigcup_{i=1}^{\ell} I_i$ and let $\delta_2 = \min\{|\bar{q}(x)|: x \in J\} > 0$.

Choose k so large that $|q_k(x)| \geq \frac{\delta_2}{2}$ for all $x \in J$ and $c_k \leq \frac{1}{2}\delta_2 m$ where $m = \min\{f(x): x \in X\}$. By Lemma 1, no two points in Y_k lie in the same I_i; furthermore, for all $x \in J$, we have

$$\left|\frac{1}{p_k(x)}\right| = \left|\frac{c_k}{q_k(x)}\right| \leq \frac{1}{2}\delta_2 m \cdot \frac{2}{\delta_2} = m.$$

Now let x_{j-1}^k, x_j^k and x_{j+1}^k be consecutive points of Y_k and suppose that $f(x_{j-1}^k) - \frac{1}{p(x_{j-1}^k)} = e_k$. Then x_j^k must lie in some I_i (since $\frac{1}{p_k(x_j^k)} = f(x_j^k) + e_k > m$) which is separated from both x_{j-1}^k and x_{j+1}^k by points of J. Since $\left|\frac{1}{p_k(x)}\right| < \frac{1}{p_k(x_j^k)}$ at such separation points, if $p_k > 0$ throughout $[x_{j-1}^k, x_{j+1}^k]$, then p_k must have a relative minimum somewhere in (x_{j-1}^k, x_{j+1}^k). If, on the other hand, $p_k(x) < 0$ for some $x \in [x_{j-1}^k, x_{j+1}^k]$ then, since $p_k(x_{j-1}^k) > 0$ and $p_k(x_{j+1}^k) > 0$, it again follows that p_k has a relative minimum somewhere in (x_{j-1}^k, x_{j+1}^k).

Next, assume x_{j-1}^k, x_j^k and x_{j+1}^k are consecutive points of Y_k with $\frac{1}{p_k(x_{j-1}^k)} - f(x_{j-1}^k) = e_k$; then $\frac{1}{p_k(x_{j-1}^k)} = f(x_{j-1}^k) + e_k > m$ implying that $x_{j-1}^k \in I_i$ for some i, and similarly for x_{j+1}^k. By Lemma 1, x_{j-1}^k and x_{j+1}^k are in distinct I_i's. If $p_k > 0$ throughout either $[x_{j-1}^k, x_j^k]$ or $[x_j^k, x_{j+1}^k]$, then for some point x in one of these intervals we have $0 < \frac{1}{p_k(x)} < \frac{1}{p_k(x_{j-1}^k)}$ and $0 < \frac{1}{p_k(x)} < \frac{1}{p_k(x_{j+1}^k)}$ since there are points of J between x_{j-1}^k and x_j^k and also between x_j^k and x_{j+1}^k.

Therefore, p_k must have a relative maximum somewhere in (x_{j-1}^k, x_{j+1}^k). If, on the other hand, $p_k < 0$ somewhere in both $[x_{j-1}^k, x_j^k]$ and $[x_j^k, x_{j+1}^k]$, then since $p_k(x_j^k) > 0$ it again follows that p_k has a relative maximum somewhere in (x_{j-1}^k, x_{j+1}^k).

We have now shown that p_k has a relative minimum between every pair of "lower extrema" of $f - \frac{1}{p_k}$ (on Y_k) and a relative maximum between every pair of "upper extrema". Thus, p_k has at least n relative extrema. But p_k is a non-trivial polynomial of degree $\leq n$. This contradiction completes the proof of the lemma.

<u>Corollary</u>. There exists a constant $c^* > 0$ such that $|p_k(x)| \leq c^*$ for $k = 1, 2, \ldots$ and all $x \in X$.

Before proceeding to Lemma 4, we introduce some new notation and make a few remarks. We shall call the exchange from Y_k to X_{k+1} an augmented exchange if $X_{k+1} = Y_k \cup \{y\}$ (recall that $Y_k \subset X_k$ is a set of $n + 2$ points on which $f - r_k$ alternates with error e_k). Also, in this case the point $y \in X$ satisfies $p_k(y) = \min\{p_k(x): \ x \in X\} < \varepsilon$. Writing $X_{k+1} = \{y_0^k, \ldots, y_{n+2}^k\}$, we have that X_{k+1} contains exactly one g-pole of r_k. Call this point y_σ^k. As stated earlier, we let r_{k+1} denote the best approximation to f from $R_n^0(X_{k+1})$ on X_{k+1} (found via the differential correction algorithm) and we define Y_{k+1} to be that subset of X_{k+1} on which $f - r_{k+1}$ alternates in sign with modulus e_{k+1}. Note that since we are assuming that we are using Exchange I, Y_{k+1} is uniquely determined by the fact that $y_\sigma^k \in Y_{k+1}$ and at precisely one point of Y_k, say t, $|f(t) - r_{k+1}(t)| < e_k$ must hold. This follows from Lemma 2. For $e_k \leq \lambda \leq e_{k+1}$ construct $\bar{r}_\lambda = \frac{1}{\bar{p}_\lambda}$ by requiring that

$$f(\tilde{y}_i^k) - \bar{r}_\lambda(\tilde{y}_i^k) = \eta_i\lambda, \ i = 0, 1, \ldots, n$$

where $\tilde{X}_{k+1} = Y_{k+1} \sim \{\tilde{y}_\sigma^k\} = \{\tilde{y}_0^k, \ldots, \tilde{y}_n^k\}$ and $\eta_i = \text{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)]$. Observe that for $\eta_i = -1$, $f(\tilde{y}_i^k) - \eta_i\lambda = f(\tilde{y}_i^k) + \lambda > 0$ and for $\eta_i = +1$, $f(\tilde{y}_i^k) - \eta_i\lambda = f(\tilde{y}_i^k) - \lambda \geq f(\tilde{y}_i^k) - e_{k+1} = r_{k+1}(\tilde{y}_i^k) \geq \frac{1}{C^*} > 0$. Thus, $\bar{p}_\lambda$ is well defined by these equations and for all $x \in X$, $\bar{p}_\lambda(x)$ is a continuous function of λ. Finally, let $\Delta = \inf\{\|f - r\| : r \in R_n^0(X)\}$. Note that $\Delta > 0$ since $f \in C(X) \sim R_n^0(X)$. Then,

<u>LEMMA 4</u>. If at the k-th exchange an augmented exchange occurs and $\text{sgn}[f(\tilde{y}_i^k) - r_k(\tilde{y}_i^k)] = \text{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)]$, $i = 0, 1, \ldots, n$, then $e_{k+1} - e_k \geq \Omega(\|f\| - \Delta)$, where Ω is a constant independent of k.

<u>Proof</u>: First observe that if $\lambda = e_k$, then $\bar{r}_{e_k} \equiv r_k$ as these two functions take on the same values on $\tilde{X}_{k+1}$ ($n + 1$ points) and likewise $\bar{r}_{e_{k+1}} \equiv r_{k+1}$. Thus, $\bar{p}_{e_k}(y_\sigma^k) < \varepsilon$ since y_σ^k is a g-pole of r_k.

We now claim that if $\tilde{y}_i^k < y_\sigma^k < \tilde{y}_{i+1}^k$ then we must have that

$$\text{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)] = \text{sgn}[f(\tilde{y}_{i+1}^k) - r_{k+1}(\tilde{y}_{i+1}^k)] = 1$$

since otherwise (i.e., = -1) a zero counting argument implies that $r_k \equiv r_{k+1}$. Similarly, if $y_\sigma^k < \tilde{y}_0^k$ or $\tilde{y}_n^k < y_\sigma^k$, we must have $\text{sgn}[f(\tilde{y}_0^k) - r_{k+1}(\tilde{y}_0^k)] = 1$ or $\text{sgn}[f(\tilde{y}_n^k) - r_{k+1}(\tilde{y}_n^k)] = 1$, respectively. This follows by counting zeros of $p_{k+1} - p_k$ in the contrary case and using the fact that $e_k < e_{k+1}$. Indeed, suppose $\tilde{y}_n^k < y_\sigma^k$ and $\text{sgn}[f(\tilde{y}_n^k) - r_{k+1}(\tilde{y}_n^k)] = -1$. Now we have that r_{k+1} alternates on $\tilde{y}_0^k, \ldots, \tilde{y}_n^k, y_\sigma^k$ with error e_{k+1} and r_k alternates on $\tilde{y}_0^k, \ldots, \tilde{y}_n^k$ with error e_k and the same sign as that of $f - r_{k+1}$. Thus, $p_{k+1} - p_k$ has n zeros in $[\tilde{y}_0^k, \tilde{y}_n^k]$. Also, we must have $r_{k+1}(\tilde{y}_n^k) > r_k(\tilde{y}_n^k)$ since we are assuming that $\text{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)] = -1$ and $e_k < e_{k+1}$. Thus, $p_{k+1}(\tilde{y}_n^k) < p_k(\tilde{y}_n^k)$. Also, since r_{k+1} is the best approximation to f on X_{k+1} we must have $p_{k+1}(y_\sigma^k) \geq \varepsilon$. Now, by construction, y_σ^k was chosen so that $p_k(y_\sigma^k) \leq \varepsilon$. Thus, $p_{k+1} - p_k$ has (at least) one additional zero in $(\tilde{y}_n^k, y_\sigma^k)$ implying $p_{k+1} \equiv p_k$ which is a contradiction. A similar argument will treat the other cases. Thus, it follows from the alternation of $f - r_{k+1}$ that $\operatorname{sgn}[f(y_\sigma^k) - r_{k+1}(y_\sigma^k)] = -1$.

Consider $\bar{p}_\lambda(y_\sigma^k)$ for $e_k \leq \lambda \leq e_{k+1}$. We know that $\bar{p}_{e_k}(y_\sigma^k) = p_k(y_\sigma^k) < \varepsilon$. Set $m = \min\{f(x): x \in X\}$ and $M = \max\{f(x): x \in X\}$. Since $\|f\| > 2\Delta > \Delta$ we have (as $\varepsilon \leq \frac{1}{2\|f\|}$)

$$
\begin{aligned}
p_{k+1}(y_\sigma^k) = \bar{p}_{e_{k+1}}(y_\sigma^k) &= (f(y_\sigma^k) + e_{k+1})^{-1} \\
&> (f(y_\sigma^k) + e_{k+1} + (\|f\| - \Delta))^{-1} \\
&> (f(y_\sigma^k) + e_{k+1} + (\|f\| - e_{k+1}))^{-1} \\
&\geq \frac{1}{2\|f\|} \geq \varepsilon > \bar{p}_{e_k}(y_\sigma^k).
\end{aligned}
$$

Therefore, $p_k(y_\sigma^k) < (f(y_\sigma^k) + e_{k+1} + (\|f\| - \Delta))^{-1} < p_{k+1}(y_\sigma^k)$. Since $\bar{p}_\lambda$ is a continuous function of λ, $e_k \leq \lambda \leq e_{k+1}$, there exists ω such that $e_k < \omega < e_{k+1}$ and

$$\bar{p}_\omega(y_\sigma^k) = (f(y_\sigma^k) + e_{k+1} + (\|f\| - \Delta))^{-1}.$$

We define coefficients $c_{j,\lambda}$ for $j = 0, 1, \ldots, n$ by setting

$$\bar{p}_\lambda(\tilde{y}_i^k) = \sum_{j=0}^{n} c_{j,\lambda}(\tilde{y}_i^k)^j$$

and let $M^* = \max_{x \in X}\{1, |x|, \ldots, |x|^n\}$. Then,

$$
\begin{aligned}
\|f\| - \Delta &= (e_{k+1} + f(y_\sigma^k) + (\|f\| - \Delta)) - (e_{k+1} + f(y_\sigma^k)) \\
&= \frac{1}{\bar{p}_\omega(y_\sigma^k)} - \frac{1}{\bar{p}_{e_{k+1}}(y_\sigma^k)} = \frac{\bar{p}_{e_{k+1}}(y_\sigma^k) - \bar{p}_\omega(y_\sigma^k)}{\bar{p}_\omega(y_\sigma^k)\bar{p}_{e_{k+1}}(y_\sigma^k)} \\
&\leq \frac{1}{\varepsilon^2}[\bar{p}_{e_{k+1}}(y_\sigma^k) - \bar{p}_\omega(y_\sigma^k)]
\end{aligned}
$$

$$\leq \frac{1}{\varepsilon^2}\left[\sum_{j=0}^{n} |c_{j,e_{k+1}} - c_{j,\omega}|\,|y_\sigma^k|^j\right]$$

$$\leq \frac{M^*(n+1)}{\varepsilon^2} \max_{0\leq j\leq n} |c_{j,e_{k+1}} - c_{j,\omega}|$$

$$= \frac{M^*(n+1)}{\varepsilon^2} |c_{\underset{\sim}{j},e_{k+1}} - c_{\underset{\sim}{j},\omega}|.$$

We complete the proof of Lemma 4 by showing that $\|f\| - \Delta \leq \Omega'(e_{k+1} - e_k)$, where Ω' is a constant independent of k. Now let

$$D(y_\sigma^k) = \det\begin{pmatrix} 1 & \tilde{y}_0^k & (\tilde{y}_0^k)^2 & \cdots & (\tilde{y}_0^k)^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \tilde{y}_n^k & (\tilde{y}_n^k)^2 & \cdots & (\tilde{y}_n^k)^n \end{pmatrix}$$

and let $D(y_\sigma^k, \lambda)$ be $D(y_\sigma^k)$ with the j-th column replaced with

$$((f(\tilde{y}_0^k) - \eta_0\lambda)^{-1}, \ldots, (f(\tilde{y}_n^k) - \eta_n\lambda)^{-1})^T$$

Also, let $\tilde{W}_i$ be the cofactor of $D(y_\sigma^k, \lambda)$ relative to the (i, j) element. Then using Cramer's rule

$$\|f\| - \Delta \leq \frac{M^*(n+1)}{\varepsilon^2}\frac{1}{|D(y_\sigma^k)|}\,|[D(y_\sigma^k, e_{k+1}) - D(y_\sigma^k, \omega)]|$$

$$= \frac{M^*(n+1)}{\varepsilon^2|D(y_\sigma^k)|}\left|\sum_{i=0}^{n} W_i\left[\frac{1}{f(\tilde{y}_i^k) - \eta_i e_{k+1}} - \frac{1}{f(\tilde{y}_i^k) - \eta_i\omega}\right]\right|$$

$$\leq \frac{M^*(n+1)}{\varepsilon^2|D(y_\sigma^k)|}\sum_{i=0}^{n} |W_i|\left[\max_{0\leq i\leq n}\frac{1}{|f(\tilde{y}_i^k) - \eta_i e_{k+1}|\,|f(\tilde{y}_i^k) - \eta_i\omega|}\right](e_{k+1} - \omega).$$

Now, since the points $\tilde{y}_i^k$ are separated (Lemma 1), there exists a constant $\xi > 0$ (independent of k) such that the Vandermonde determinant $|D(y_\sigma^k)| \geq \xi > 0$. Furthermore, $\sum_{i=0}^{n} |W_i| \leq K$ for some positive constant K since all cofactors of $D(y_\sigma^k)$ are bounded (independent of k). Finally,

$$|f(\tilde{y}_i^k) - \eta_i e_{k+1}| = \frac{1}{\bar{p}_{e_{k+1}}(\tilde{y}_i^k)} = \frac{1}{p_{k+1}(\tilde{y}_i^k)} \geq \frac{1}{e^*}$$

by the corollary following Lemma 3. Furthermore, $|f(\tilde{y}_i^k) - \eta_i \omega| = \dfrac{1}{\bar{p}_\omega(\tilde{y}_i^k)}$, and by construction this lies between $f(\tilde{y}_i^k)$ and $\dfrac{1}{\bar{p}_{e_{k+1}}(\tilde{y}_i^k)}$, so that

$$|f(\tilde{y}_i^k) - \eta_i \omega| \geq \min(m, \frac{1}{e^*}).$$

Thus,

$$\max_{0 \leq i \leq n} \frac{1}{|f(\tilde{y}_i^k) - \eta_i e_{k+1}|\ |f(\tilde{y}_i^k) - \eta_i \omega|} \leq e^* \max(e^*, \frac{1}{m}).$$

Therefore, $\|f\| - \Delta \leq \Omega'(e_{k+1} - \omega)$. Taking $\Omega = \frac{1}{\Omega'}$ yields

$$\Omega(\|f\| - \Delta) \leq e_{k+1} - \omega \leq e_{k+1} - e_k. \qquad \blacksquare$$

Now let us assume there exists a subsequence of positive integers $\{k_m\}_{m=1}^{\infty}$ satisfying the following:

1. An augmented exchange occurs between X_{k_m} and X_{k_m+1}

2. $e_{k_m+1} - e_{k_m} < \Omega(\|f\| - \Omega)$ (since $e_{k_m} \uparrow e^* \leq \Delta$, where Δ is the error of best approximation to f from $R_n^0(X)$).

By our assumption 2, we see that the sign condition of Lemma 4 cannot hold, hence we have for each k_m, the additional condition:

3. $\text{sgn}[f(\tilde{y}_i^{k_m}) - r_{k_m+1}(\tilde{y}_i^{k_m})] \neq \text{sgn}[f(\tilde{y}_i^{k_m}) - r_{k_m}(\tilde{y}_i^{k_m})]$ for some i, $0 \leq i \leq n$.

Recall that for an augmented exchange between X_{k_m} and X_{k_m+1}, that Y_{k_m+1} denotes the subset of X_{k_m+1} consisting of n + 2 points on which $f - r_{k_m+1}$ alternates with error e_{k_m+1}. Define $y_\beta^{k_m}$ by $\{y_\beta^{k_m}\} = Y_{k_m} \sim Y_{k_m+1}$. That is, $y_\beta^{k_m}$ is the point of X_{k_m+1} which is deleted in forming Y_{k_m+1}. Since we are considering

Exchange procedure I, we have that W_{k_m} may be taken to be Y_{k_m} whenever an augmented exchange is performed. Under these assumptions we prove the following two lemmas.

LEMMA 5. If $\bar{p}_{e_k}(y_\beta^k) < \frac{\varepsilon}{(1+\varepsilon)}$, then $e_{k+1} - e_k \geq \Omega''$ where Ω'' is a constant independent of k.

Proof: Once again, define $\bar{r}_\lambda = \frac{1}{\bar{p}_\lambda}$, $e_k \leq \lambda \leq e_{k+1}$ by

$$\bar{p}_\lambda(\tilde{y}_i^k) = (f(\tilde{y}_i^k) - \eta_i e_k)^{-1}, \; 0 \leq i \leq n$$

with $\eta_i = \mathrm{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)]$, $0 \leq i \leq n$. Since we are no longer assuming that $\mathrm{sgn}[f(\tilde{y}_i^k) - r_k(\tilde{y}_i^k)] = \mathrm{sgn}[f(\tilde{y}_i^k) - r_{k+1}(\tilde{y}_i^k)]$ for all i, we do not necessarily have that $\bar{p}_{e_k} \equiv p_k$. However, we still have that $\bar{p}_{e_{k+1}} \equiv p_{k+1}$ as before so that $\bar{p}_{e_{k+1}}(y_\beta^k) \geq \varepsilon > \frac{\varepsilon}{(1+\varepsilon)}$ holds. Since $\bar{p}_{e_k}(y_\beta^k) < \frac{\varepsilon}{(1+\varepsilon)}$ by hypothesis we have, by the Intermediate Value Theorem, that there exists ω, $e_k < \omega < e_{k+1}$, such that $\bar{p}_\omega(y_\beta^k) = \frac{\varepsilon}{(1+\varepsilon)}$. Then, as before

$$1 = \frac{1+\varepsilon}{\varepsilon} - \frac{1}{\varepsilon} \leq \frac{1}{\bar{p}_\omega(y_\beta^k)} - \frac{1}{\bar{p}_{e_{k+1}}(y_\beta^k)} = \frac{\bar{p}_{e_{k+1}}(y_\beta^k) - \bar{p}_\omega(y_\beta^k)}{\bar{p}_\omega(y_\beta^k)\bar{p}_{e_{k+1}}(y_\beta^k)}$$

$$\leq \frac{1+\varepsilon}{\varepsilon^2} \sum_{j=0}^{n} (c_{j,e_{k+1}} - c_{j,\omega})(y_\beta^k)^j$$

where the coefficients $c_{j,e_{k+1}}$ and $c_{j,\omega}$ are as defined in the proof of Lemma 4. Since the estimates used in the proof of Lemma 4 are independent of the point y_σ^k, they may be applied here for y_β^k, and we have $1 \leq \Omega_0(e_{k+1} - \omega) \leq \Omega_0(e_{k+1} - e_k)$ implying $e_{k+1} - e_k \geq \frac{1}{\Omega_0} = \Omega'' > 0$.

LEMMA 6. An augmented exchange can occur only a finite number of times.

Proof: Assume the contrary. Then, by extracting subsequences (as often as necessary) we obtain a sequence $\{y_{k_\ell}\}$ of extreme points $\{x_0^{k_\ell}, \ldots, x_{n+1}^{k_\ell}\}$ such that the exchange from Y_{k_ℓ} to $X_{k_\ell+1}$ is an augmented exchange, so that $X_{k_\ell+1} = Y_{k_\ell} \cup \{y_\sigma^{k_\ell}\}$ where $y_\sigma^{k_\ell}$ is a g-pole for r_{k_ℓ} and is selected so that $p_k(y_\sigma^{k_\ell}) = \min\{p_{k_\ell}(y)\colon\ y \in X\} < \varepsilon$; r_{k_ℓ} is the best approximation to f on $Z_{k_\ell} = \{x \in X\colon\ p_{k_\ell}(x) \geq \varepsilon\}$ from $R_n^0(Z_{k_\ell})$ so that no multiple exchange is applied to Y_{k_ℓ}. Letting $y_\sigma^{k_\ell}$ be defined as before, $\{y_\beta^{k_\ell}\} = Y_{k_\ell} \sim Y_{k_\ell+1}$, we assume that the final subsequence $\{Y_{k_\ell}\}$ for k = 1, 2, ... satisfies

(1) $\bar{p}_{e_{k_\ell}}(y_\beta^{k_\ell}) \geq \dfrac{\varepsilon}{(1+\varepsilon)}$ (Lemma 5).

(2) $\operatorname{sgn}[f(\tilde{y}_i^{k_\ell}) - r_{k_\ell}(\tilde{y}_i^{k_\ell})] \neq \operatorname{sgn}[f(\tilde{y}_i^{k_\ell}) - r_{k_\ell+1}(\tilde{y}_i^{k_\ell})]$ for some i, $0 \leq i \leq n$, independent of k_ℓ, where $X_{k_\ell+1} = Y_{k_\ell+1} - \{y_\sigma^{k_\ell}\} = \{y_0^{k_\ell}, \ldots, y_n^{k_\ell}\}$ (Lemma 4).

(3) $y_\beta^{k_\ell} \to y_\beta^* \in X$ as $k_\ell \to \infty$

(4) $\tilde{X}_{k_\ell+1} \to X^* = \{y_0^*, \ldots, y_n^*\} \subset X$ (coordinatewise convergence) with $y_{j+1}^* - y_j^* \geq \delta > 0$, $0 \leq j \leq n$ (Lemma 1).

(5) $p_{k_\ell} \to p^* \in \Pi_n$ uniformly on X (Lemma 2).

(6) $\operatorname{sgn}[f(\tilde{y}_j^{k_\ell}) - r_k(\tilde{y}_j^{k_\ell})]$ is constant for fixed j, independent of k_ℓ.

As noted above, all of these conditions can be met by passing to subsequences of subsequences sufficiently often. Now, under these conditions, we claim there exists a $\rho > 0$, independent of k_ℓ, such that

$$|f(y_\beta^{k_\ell}) - \bar{r}_{e_{k_\ell}}(y_\beta^{k_\ell})| \geq e_{k_\ell} + \rho \tag{2.1}$$

where $\bar{r}_\lambda$ is defined for $e_{k_\ell} \leq \lambda \leq e_{k_\ell+1}$ as before:

$$\bar{r}_\lambda(\tilde{y}_j^{k_\ell}) = f(\tilde{y}_j^{k_\ell}) - \eta_j\lambda, \qquad 0 \leq j \leq n$$

with $\eta_j = \operatorname{sgn}[f(\tilde{y}_j^{k_\ell}) - r_{k_\ell+1}(\tilde{y}_j^{k_\ell})]$. Indeed, if (2.1) is not true, then there exists a subsequence (relabelled) such that

$$|f(y_\beta^{k_\ell}) - \bar{r}_{e_{k_\ell}}(y_\beta^{k_\ell})| \leq e_{k_\ell} + \frac{1}{k_\ell}. \tag{2.2}$$

Define $q \in \Pi_n$ by

$$q(y_j^*) = \lim_{k_\ell\to\infty} (f(\tilde{y}_j^{k_\ell}) - \eta_j e_{k_\ell})^{-1} = \lim_{k_\ell\to\infty} \bar{p}_{e_{k_\ell}}(\tilde{y}_j^{k_\ell}),\ 0 \leq j \leq n. \tag{2.3}$$

Note that if $\eta_j = -1$, then $f(\tilde{y}_j^{k_\ell}) - \eta_j e_{k_\ell} \geq f(\tilde{y}_j^{k_\ell}) \geq \min\{f(x): x \in X\} = m > 0$, and if $\eta_j = 1$, then $f(\tilde{y}_j^{k_\ell}) - \eta_j e_{k_\ell} \geq f(\tilde{y}_j^{k_\ell}) - e_{k_\ell+1} = r_{k_\ell+1}(\tilde{y}_j^{k_\ell}) \geq \frac{1}{c^*} > 0$. Thus, the above limit exists for each j and $q(x) > 0$ on X^*. From this it follows that $\bar{p}_{e_k}$ converges uniformly to q on X. Furthermore, $q(y_\beta^*) \geq \frac{\varepsilon}{(1+\varepsilon)}$ by (1) and by (2.3) $|f(y_j^*) - \frac{1}{q(y_j^*)}| = e^*$. Thus, for k_ℓ sufficiently large, say $k_\ell \geq \bar{k}$, so that $q(y_\beta^{k_\ell}) > 0$ we have

$$|f(y_\beta^*) - \frac{1}{q(y_\beta^*)}| \leq |f(y_\beta^*) - f(y_\beta^{k_\ell})| + |f(y_\beta^{k_\ell}) - \bar{r}_{e_{k_\ell}}(y_\beta^{k_\ell})|$$
$$+ |\bar{r}_{e_{k_\ell}}(y_\beta^{k_\ell}) - \frac{1}{q(y_\beta^{k_\ell})}| + |\frac{1}{q(y_\beta^{k_\ell})} - \frac{1}{q(y_\beta^*)}| \to e^* \text{ as } k_\ell \to \infty.$$

Now, since $p_{k_\ell}(\tilde{y}_j^{k_\ell}) \geq \varepsilon$ for all j and k_ℓ and $p_{k_\ell}(y_\beta^{k_\ell}) \geq \varepsilon$ for all k_ℓ, we have that $p^*(y_j^*) \geq \varepsilon$ for all j and $p^*(y_\beta^*) \geq \varepsilon$. Furthermore, $f - \frac{1}{p^*}$ alternates on

$X^* \cup \{y_\beta^*\}$ with deviation e^* since $f - \frac{1}{p_{k_\ell}}$ alternates on $\tilde{X}_{k_\ell+1} \cup \{y_\beta^{k_\ell}\}$ with deviation e_{k_ℓ}. Thus, by Lemma 2, $\frac{1}{q} = \frac{1}{p^*}$. But this is impossible since for k_ℓ sufficiently large (i the index of (2))

$$\operatorname{sgn}[f(y_j^*) - \frac{1}{p^*(y_i)}] = \operatorname{sgn}[f(\tilde{y}_i^{k_\ell}) - r_{k_\ell}(\tilde{y}_i^{k_\ell})]$$

$$= -\operatorname{sgn}[f(\tilde{y}_i^{k_\ell}) - r_{k_\ell+1}(\tilde{y}_i^{k_\ell})] = -\operatorname{sgn}[f(\tilde{y}_i^{k_\ell}) - \bar{r}_{e_{k_\ell+1}}(\tilde{y}_i^{k_\ell})]$$

$$= -\operatorname{sgn}[f(y_i^*) - \frac{1}{q(y_i^*)}] \neq 0,$$

which is our desired contradiction. Thus, (2.1) holds. Since $e_{k_\ell+1} > e_{k_\ell} \uparrow e^*$, we have that $e_{k_\ell+1} - e_{k_\ell} \to 0$. Choose $\tilde{k}$ so that $k_\ell \geq \tilde{k}$ implies $e_{k_\ell} + \frac{\rho}{2} > e_{k_\ell+1}$. Then, we have, using (2.1) and the determinant argument from the proof of Lemma 4,

$$\frac{\rho}{2} < e_{k_\ell} + \rho - e_{k_\ell+1} \leq |f(y_\beta^{k_\ell}) - \bar{r}_{e_{k_\ell}}(y_\beta^{k_\ell})| - |f(y_\beta^{k_\ell}) - \bar{r}_{e_{k_\ell+1}}(y_\beta^{k_\ell})|$$

$$\leq \left|\frac{1}{\bar{p}_{e_{k_\ell}}(y_\beta^{k_\ell})} - \frac{1}{\bar{p}_{e_{k_\ell+1}}(y_\beta^{k_\ell})}\right| = \frac{\left|\bar{p}_{e_{k_\ell+1}}(y_\beta^{k_\ell}) - \bar{p}_{e_{k_\ell}}(y_\beta^{k_\ell})\right|}{\bar{p}_{e_{k_\ell}}(y_\beta^{k_\ell}) \cdot \bar{p}_{e_{k_\ell+1}}(y_\beta^{k_\ell})}$$

$$\leq \frac{1+\varepsilon}{\varepsilon^2}\left|\bar{p}_{e_{k_\ell+1}}(y_\beta^{k_\ell}) - \bar{p}_{e_{k_\ell}}(y_\beta^{k_\ell})\right| \leq \Omega_1(e_{k_\ell+1} - e_{k_\ell}).$$

Thus, $e_{k_\ell+1} - e_{k_\ell} \geq \Omega_2$, $\Omega_2 = \frac{\rho}{2\Omega_1}$. But this is impossible, so we have that an augmented exchange can occur only a finite number of times. ∎

We now turn our attention to the case that the exchange from Y_k to X_{k+1} is not an augmented exchange. In this case, r_k is not the best approximation to f on Z_k from $R_n^0(Z_k)$ and $X_{k+1} = Y_{k+1} = \{x_0^{k+1}, \ldots, x_{n+1}^{k+1}\}$ with no g-pole of r_k in

X_{k+1}. Setting $\gamma_{k+1} = \min_{0 \leq i \leq n+1} |f(x_i^{k+1}) - r_k(x_i^{k+1})|$ and $\beta_{k+1} = \max_{0 \leq i \leq n+1} |f(x_i^{k+1}) - r_k(x_i^{k+1})|$ we observe that $\gamma_{k+1} \geq e_k$ and $\beta_{k+1} > e_{k+1}$.

<u>LEMMA 7</u>. There exists a constant $\Omega > 0$ (independent of k) such that if X_{k+1} is not obtained by an augmented exchange then $e_{k+1} - \gamma_{k+1} \geq \Omega(\beta_{k+1} - e_{k+1})$.

<u>Proof</u>: Let λ be a parameter satisfying $\gamma_{k+1} \leq \lambda \leq e_{k+1}$. Set $\eta = \operatorname{sgn}[f(x_0^{k+1}) - r_k(x_0^{k+1})]$ and note that $(-1)^i\eta = \operatorname{sgn}[f(x_i^{k+1}) - r_k(x_i^{k+1})]$ for $i = 0, 1, \ldots, n+1$. In addition, it is always true that $\operatorname{sgn}[f(x_i^{k+1}) - r_k(x_i^{k+1})] = \operatorname{sgn}[f(x_i^{k+1}) - r_{k+1}(x_i^{k+1})]$. This fact follows from a zero counting argument since both r_k and r_{k+1} are positive on Y_{k+1} and both $f - r_k$ and $f - r_{k+1}$ alternate in sign on Y_{k+1}. If $(-1)^i\eta = 1$, then we have that $f(x_i^{k+1}) - (-1)^i\eta\lambda = f(x_i^{k+1}) - \lambda \geq f(x_i^{k+1}) - e_{k+1} = r_{k+1}(x_i^{k+1}) \geq c^* > 0$. On the other hand, if $(-1)^i\eta = -1$, we have that $f(x_i^{k+1}) - (-1)^i\eta\lambda = f(x_i^{k+1}) + \lambda \geq f(x_i^{k+1}) \geq m = \min\{f(x): x \in X\} > 0$. In either case we have that

$$f(x_i^{k+1}) - (-1)^i\eta\lambda > 0 \text{ for } 0 \leq i \leq n+1, \ \gamma_{k+1} \leq \lambda \leq e_{k+1}.$$

Define $\bar{p}_\lambda \in \Pi_n$ by $\bar{p}_\lambda(x_i^{k+1}) = (f(x_i^{k+1}) - (-1)^i\eta\lambda)^{-1}$ for $i = 0, 1, \ldots, n+1$, $i \neq q$, where q is the smallest subscript, $0 \leq q \leq n+1$, for which $|f(x_q^{k+1}) - r_k(x_q^{k+1})| = \beta_{k+1}$. Next, define $\bar{r}_\lambda = \frac{1}{\bar{p}_\lambda}$ and note that $\bar{r}_{e_{k+1}} \equiv r_{k+1}$ since these two functions agree at n + 1 points (i.e., x_i^{k+1}, $i = 0, 1, \ldots, n+1$, $i \neq q$). Finally, observe that $\bar{p}_\lambda(x_q^{k+1})$ is a continuous function of λ for $\gamma_{k+1} \leq \lambda \leq e_{k+1}$.

We shall prove that there exists an ω, $\gamma_{k+1} \leq \omega < e_{k+1}$ such that $\bar{p}_\omega(x_q^{k+1}) = p_k(x_q^{k+1})$. To do this we must consider two cases:

Case 1: $(-1)^q\eta = 1$, i.e., $\mathrm{sgn}[f(x_q^{k+1}) - r_{k+1}(x_q^{k+1})] = 1$. Here $f(x_q^{k+1}) - r_k(x_q^{k+1}) = \beta_{k+1}$ and $f(x_q^{k+1}) - r_{k+1}(x_q^{k+1}) = e_{k+1}$. For $i \neq q$, we have $|f(x_i^{k+1}) - \bar{r}_{\gamma_{k+1}}(x_i^{k+1})| = \gamma_{k+1} \leq |f(x_i^{k+1}) - r_k(x_i^{k+1})|$, so that $(-1)^i\eta[r_k(x_i^{k+1}) - \bar{p}_{\gamma_{k+1}}(x_i^{k+1})] \leq 0$, and thus $(-1)^i\eta[p_k(x_i^{k+1}) - \bar{p}_{\gamma_{k+1}}(x_i^{k+1}) \geq 0$. $i \neq q$. Now, if $(-1)^q\eta[p_k(x_q^{k+1}) - \bar{p}_{\gamma_{k+1}}(x_q^{k+1}) \geq 0$ holds, then by counting zeros (including multiplicities of up to order 2) one has that $p_k \equiv \bar{p}_{\gamma_{k+1}}$, so that $p_k(x_q^{k+1}) = \bar{p}_{\gamma_{k+1}}(x_q^{k+1})$ and one sets $\omega = \gamma_{k+1}$ in this case. If, on the other hand $(-1)^q\eta[p_k(x_q^{k+1}) - \bar{p}_{\gamma_{k+1}}(x_q^{k+1})] < 0$ holds, then $\bar{p}_{\gamma_{k+1}}(x_q^{k+1}) > p_k(x_q^{k+1})$. Since $f(x_q^{k+1}) - r_k(x_q^{k+1}) = \beta_{k+1} > e_{k+1} = f(x_q^{k+1}) - r_{k+1}(x_q^{k+1})$ we also have that $p_k(x_q^{k+1}) > p_{k+1}(x_q^{k+1}) = \bar{p}_{e_{k+1}}(x_q^{k+1})$ so that by the Intermediate Value Theorem there is an ω, $\gamma_k < \omega < e_{k+1}$ such that $\bar{p}_\omega(x_q^{k+1}) = p_k(x_q^{k+1})$.

Case 2: $(-1)\eta = -1$. This case follows with essentially the same argument and we shall not give the details.

Thus, there exists an ω, $\gamma_{k+1} \leq \omega < e_{k+1}$ such that $\bar{p}_\omega(x_q^{k+1}) = p_k(x_q^{k+1})$.

Hence

$$\beta_{k+1}-e_{k+1}=f(x_q^{k+1})-\bar{r}_\omega(x_q^{k+1})-[f(x_q^{k+1})-r_{e_{k+1}}(x_q^{k+1})]=\frac{1}{\bar{p}_{e_{k+1}}(x_q^{k+1})} - \frac{1}{\bar{p}_\omega(x_q^{k+1})}$$

$$=\frac{1}{\bar{p}_\omega(x_q^{k+1})\bar{p}_{e_{k+1}}(x_q^{k+1})} \sum_{j=0}^{n} [c_{j,\omega} - c_{j,e_{k+1}}](x_q^{k+1})^j.$$

Since $\bar{p}_\omega(x_q^{k+1}) = p_k(x_q^{k+1}) \geq \varepsilon$, and referring to the already established estimates of Lemma 4, we have

$$\beta_{k+1} - e_{k+1} \leq \frac{M^*(n+1)}{\varepsilon^2|D(x_q^{k+1})|} \sum_{\substack{i=0 \\ i\neq q}}^{n} |W_i|[\max(c^*, \frac{1}{m})](e_{k+1} - \omega)$$

$$\leq \Omega'(e_{k+1} - \omega) \leq \Omega'(e_{k+1} - \gamma_{k+1}),$$

where $\Omega' > 0$ is independent of k. Setting $\Omega = \frac{1}{\Omega'}$ we have our desired result. ∎

Finally, collecting all the above results to complete the proof of Theorem 2, we see that first of all there exists a positive integer k_0 such that for $k \geq k_0$ no augmented exchanges occur. Thus, for all $k \geq k_0$ we have that $e_{k+1} - \gamma_{k+1} \geq \Omega(\beta_{k+1} - e_{k+1}) = \Omega(\beta_{k+1} - \gamma_{k+1}) - \Omega(e_{k+1} - \gamma_{k+1})$, by Lemma 7. Hence, for $k \geq k_0$, $e_{k+1} - \gamma_{k+1} \geq [\frac{\Omega}{(1 + \Omega)}](\beta_{k+1} - \gamma_{k+1})$ implying that $\gamma_{k+2} - \gamma_{k+1} \geq [\frac{\Omega}{(1 + \Omega)}](\beta_{k+1} - \gamma_{k+1})$.

Now we may apply the argument given in the continuous case ([2], p. 99), noting that we have a Strong Uniquenss Theorem ([5, Theorem 3]), to show that there exists $\theta \in (0, 1)$ and $A > 0$ such that if $k > k_0$ then

$$\|f - r_k\| \leq A\theta^k$$

completes the proof of Theorem 2.

3. Approximation from $R_n^m(X)$

We now turn to the second objective of this paper. Here our approximating family is taken to be

$$R_n^m(X) = \{r = \frac{p}{q}: \ p \in \Pi_n, \ q \in \Pi_n, \ q > 0 \text{ on } X\},$$

and require card(X) $\geq m + n + 2$ $(m \geq 0,\ n \geq 0)$. g-poles are defined as before, i.e., $x \in X$ is said to be a g-pole of $r = \frac{p}{q}$ if $q(x) < \varepsilon$ where $\varepsilon > 0$. This concept is useful even when X is finite, since it enables us to avoid division by very small positive numbers. We have used $\varepsilon = 10^{-18}$ on a UNIVAC 1106, which has roughly 18-digit accuracy in double precision. Unfortunately, we can no longer be sure that r_k will be g-pole free on its reference set, although this condition can be enforced by inserting additional constraints into the linear programming part of the differential correction algorithm (we will return to this point later). The algorithm we used (with card(X) = NUMGR < ∞) is described by the flowcharts 1 and 2.

4. Convergence of the Remes-Difcor algorithm

In this section we prove that if the 20-step stopping criterion is deleted from the Remes-Difcor flow chart, then under certain existence assumptions the algorithm will terminate at a best approximation to f from $R_n^m(X)$.

THEOREM 3. Let X be a finite set of real numbers containing at least $m + n + 3$ points, and let $f \in C(X)$. Suppose that for each subset $Y \subset X$ containing exactly $m + n + 2$ or $m + n + 3$ points, a best approximation $r = \frac{p}{q} \in R_n^m(Y)$ exists for f from $R_n^m(Y)$ and, in addition, that $q \geq \varepsilon$ on Y. Then the Remes-Difcor algorithm will terminate at a best approximation r^* to f on X from $R_n^m(X)$.

Proof: Let X_0 be the initial reference set and let X_k be the reference set at the k-th stage. Let r_k be the best approximation to f on X_k with $e_k = \max\{|f(x) - r_k(x)|: x \in X_k\}$. If the algorithm terminates at stage k, then there are no g-poles and the maximum error occurs in X_k; thus, $e_k = \|f - r_k\|$ and r_k is the best approximation to f on X from $R_n^m(X)$.

Now suppose the algorithm does not terminate at the k-th stage ($k \geq 1$). If r_{k-1} has g-poles in X, then at least one of these is included in X_k by construction, so that $r_k \not\equiv r_{k-1}$. Also, if r_{k-1} has no g-poles in X, then $r_k \not\equiv r_{k+1}$; since otherwise the maximum error for r_k in X would occur at some point as the maximum error for r_{k-1}, and thus would be included in X_k. This would contradict the fact that the algorithm does not terminate at the k-th stage.

Now $f - r_{k-1}$ must alternate on some set $\{x_1, x_2, \ldots, x_{m+n+2-d_{k-1}}\} \subset X_{k-1}$ where d_{k-1} is the defect of r_{k-1}, and so by construction $f - r_{k-1}$ must alternate in sign on some set $\{x_1', x_2', \ldots, x'_{m+n+2-d_{k-1}}\} \subset X_k$ with $|f(x_i') - r_{k-1}(x_i')| \geq e_{k-1}$,

$i = 1, 2, \ldots, m + n + 2 - d_{k-1}$. So by Lemma 2 we have $e_k = \max_{x \in X_k} |f(x) - r_k(x)|$ $\geq \max_i |f(x_i') - r_k(x_i')| > \min_i |f(x_i') - r_{k-1}(x_i')| \geq e_{k-1}$. Therefore, $\{e_k\}$ is strictly increasi[ng], so since there are only a finite number of possible reference sets containe[d in] X the algorithm must terminate. ∎

5. Examples and Conclusions

In order to get a time comparison of the Remes-Difcor algorithm with the ordinary differential correction algorithm alone, we ran the following digital filter design problem:

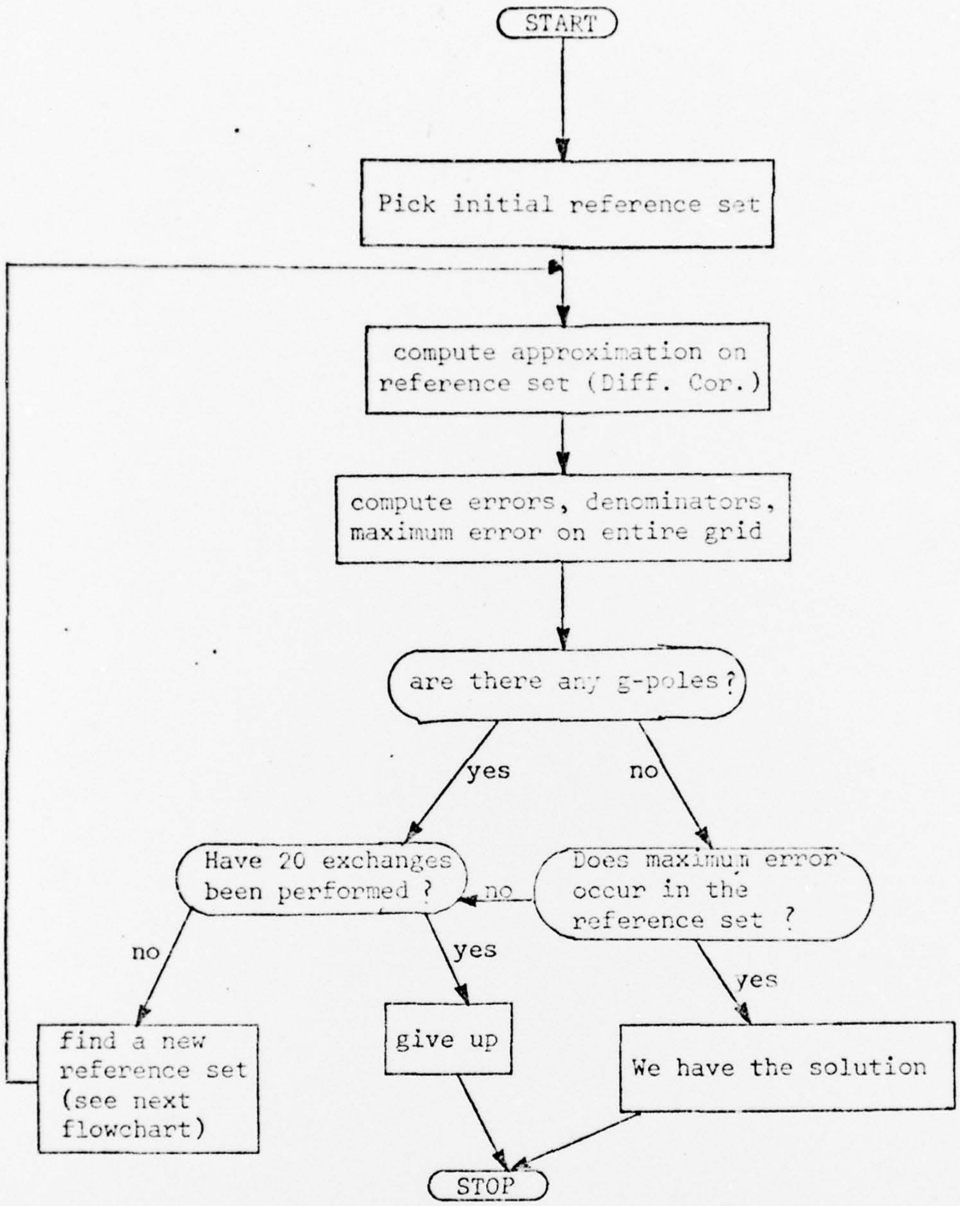
$$\text{Let } X = [0,\ 0.2\pi] \cup [0.4\pi,\ \pi],\ f(x) = \begin{cases} 1, & 0 \le x \le 0.2\pi \\ 0.0123, & 0.4\pi \le x \le \pi \end{cases}$$

We approximate from $\tilde{R}_2^9(X) = \{\frac{p(x)}{q(x)} = (a_0 + a_1 \cos x + \ldots + a_9 \cos 9x)/(b_0 + b_1 \cos x + b_2 \cos 2x) : q > 0 \text{ on } X\}$. We also want $q > 0$ on $[0, \pi]$ and $\frac{p}{q} \ge 0$ on $[0, \pi]$, but in this example it is not necessary to do anything extra to force this. Although we are not using ordinary algebraic rational functions, we do have the alternating theory in this situation, and that is all that is required. To run this example we replaced X with an equally-spaced mesh (spacing $\frac{\pi}{256}$) containing 206 points. Using as our initial reference set five (roughly) equally spaced points in $[0, 0.2\pi]$ and eight (roughly) equally spaced points in $[0.4\pi, \pi]$, we obtained convergence after four exchanges and 60.0 seconds; $\|f - r^*\|$ was 1.83914×10^{-4} (where r^* is best). (Note: the final alternating set does have five points in $[0, 0.2\pi]$ and eight points in $[0.4\pi, \pi]$, but they are not equally spaced.) Starting with eight equally spaced points in $[0, 0.2\pi]$ and five in $[0.4\pi, \pi]$, eight exchanges and 1 minute 11.7 seconds were required; starting with all reference points pushed to the extreme right of $[0.4\pi, \pi]$ (which is one of the worst possible starting reference sets), fifteen exchanges and 2 minutes 0.5 seconds were required. On the other hand, running this problem with differential correction alone required 5 minutes 45.5 seconds. One would expect the time difference to increase if a finer mesh were used.

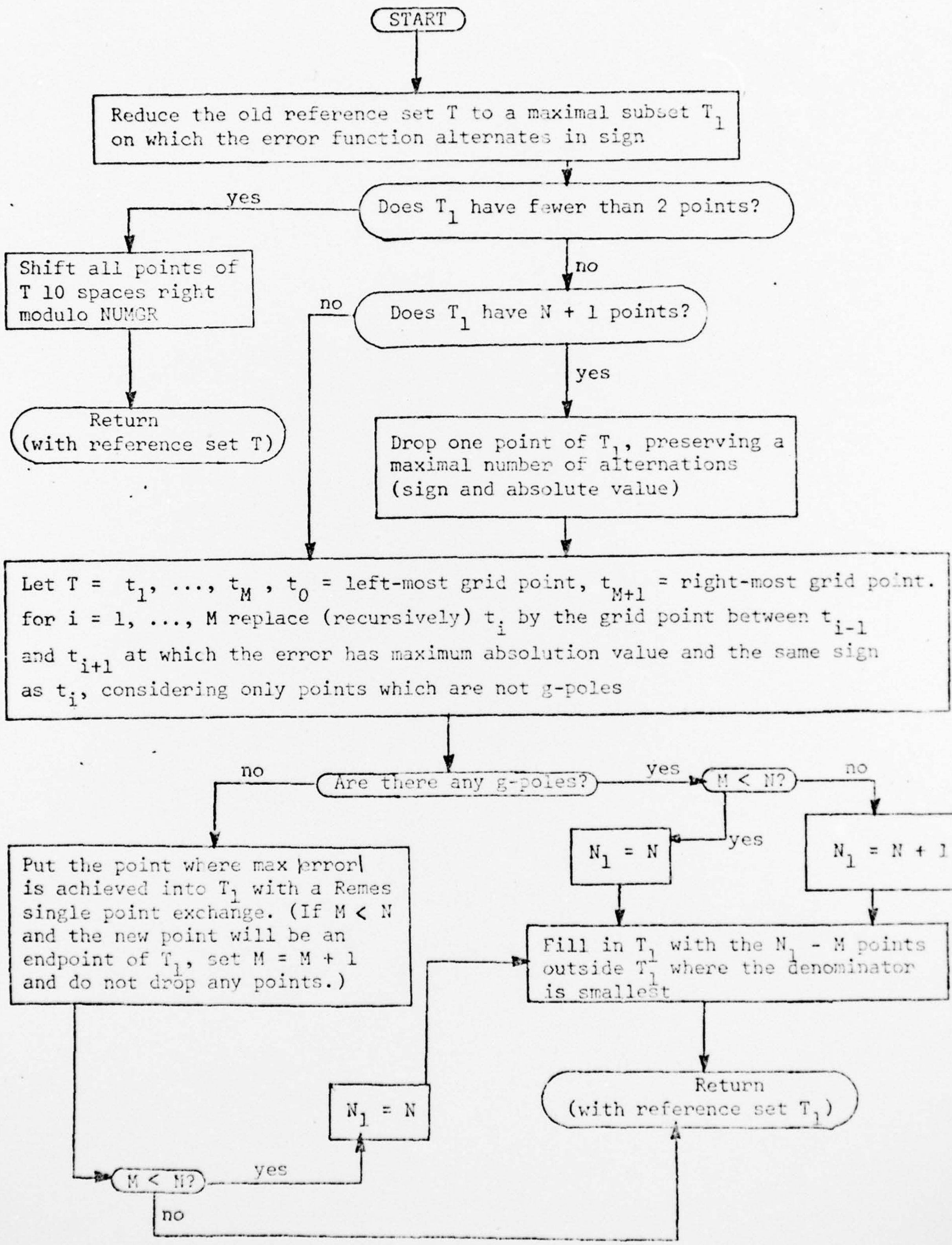
We also ran the Remes-Difcor program on an example for which best approximations did not exist on some reference sets, although a best approximation did exist on X. Here convergence depended on the choice of initial reference set, although we were able to obtain convergence even with a bad initial reference set if we "helped the program over the bad spots" by forcing $q \geq \varepsilon$ on the reference set; this (as opposed to forcing $q \geq \varepsilon$ on all of X) did not require much additional work.

In general, the relative merits of Remes, Remes-difcor, and difcor for finite X can be summarized as follows. When Remes works, so does Remes-difcor, and with comparable speed. Remes-difcor will usually still work when Remes fails due to problems in finding a new approximation on a reference set, and is much faster than difcor if card(X) is large. Difcor is theoretically more robust than Remes-difcor since it does not require an alternating theory, and $\|f - r_k\|$ will converge to $\inf_r \|f - r\|$ even if there is no best approximation, but round-off and storage problems may be prohibitive if card(X) is too large.

Remes-Difcor Flowchart #1 (excluding input-output). For a fixed number (NUMGR) of grid points.



Flowchart #2. Finding a New Reference Set ($N = n + m + 2$, NUMGR = number of grid points)



REFERENCES

1. I. Barrodale, M.J.D. Powell and F.D.K. Roberts, The differential correction algorithm for rational ℓ_∞ approximation, SIAM J. Numer. Anal. 9 (1972) 493-504.

2. E.W. Cheney, Introduction to Approximation Theory, McGraw-Hill, New York, 1966.

3. C.B. Dunham, A difficulty in Williams' algorithm for interpolating rationals, Math. Comp., 29 (1975) 552-553.

4. E.H. Kaufman, Jr. and G.D. Taylor, Uniform rational approximation by functions of several variables, Int. J. for Num. Meth. Eng. 9 (1975) 297-323.

5. D.J. Leeming and G.D. Taylor, Approximation with reciprocals of polynomials on compact sets (to appear).

6. B. Wendroff, Theoretical Numerical Analysis, Academic Press, New York, 1966.

7. J. Williams, Numerical Chebyshev approximation by interpolating rationals, Math. Comp. 26 (1972) 199-206.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFOSR - TR - 77 - 0165	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) A COMBINED REMES-DIFFERENTIAL CORRECTION ALGORITHM FOR RATIONAL APPROXIMATION	5. TYPE OF REPORT & PERIOD COVERED Interim rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Edwin H. Kaufman, Jr., David J. Leeming and G. D. Taylor	8. CONTRACT OR GRANT NUMBER(s) AFOSR-76-2878-76
9. PERFORMING ORGANIZATION NAME AND ADDRESS Colorado State University Department of Mathematics Fort Collins, Colorado 80523	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bldg. 410, Bolling AFB Washington, D.C. 20332	12. REPORT DATE November, 1976
	13. NUMBER OF PAGES 30
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 33p.	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computation of best uniform rational approximations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper a hybrid Remes-differential correction algorithm for computing best uniform rational approximants on a compact subset of the real line is developed. This algorithm differs from the classical multiple exchange Remes algorithm in two crucial aspects. First of all, the solving of a nonlinear system to find a best approximation on a given reference set in each iteration of the Remes algorithm is replaced with the differential correction algorithm to compute the desired best approximation on the reference set. →

next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

407 344

20. Secondly, the exchange procedure itself has been modified to eliminate the possibility of cycling that can occur in the usual exchange procedure. This second modification is necessary to guarantee the convergence of this algorithm on a finite set without the usual normal and sufficiently dense assumptions that exist in other studies.